УДК 004.932
В. П. Машталир, д-р техн. наук,
С. И. Богучарский

СЕГМЕНТАЦИЯ ИЗОБРАЖЕНИЙ В БОЛЬШИХ БАЗАХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ИНФОРМАЦИИ

Аннотация. *Рассмотрена задача сегментации изображений на основе алгоритма кластеризации с использованием плотности распределения информации. Особенностью предложенного метода является возможность формирования кластеров произвольной формы, обработка сигналов, зашумленных разного вида возмущениями, матричное представление информации, хранящейся в больших базах данных*

Ключевые слова: *сегментация, кластеризация, плотность распределения, обработка изображений, принадлежность*

V. Mashtalir, ScD,
S. Bogucharskiy

IMAGE SEGMENTATION IN VELDT WITH DENSITY OF INFORMATION DISTRIBUTION USAGE

Abstract. *Image segmentation based on clustering with density of information distribution usage has been considered. Proposed technique feature is a possibility of arbitrary cluster shape generation, signal processing under different disturbances, matrix representation of information stored in VLDB.*

Keywords: *segmentation, clustering, density of distribution, image processing, membership*

В. П. Машталір, д-р техн. наук,
С. І. Богучарський

СЕГМЕНТАЦІЯ ЗОБРАЖЕНЬ У ВЕЛИКИХ БАЗАХ ДАНИХ З ВИКОРИСТАННЯМ ЩІЛЬНІСТІ РОЗПОДІЛУ ІНФОРМАЦІЇ

Анотація. *Розглянуто задачу сегментації зображень на основі алгоритма кластеризації з використанням щільності розподілу інформації. Особлівістю запропонованого методу є можлівість формування кластерів довільної форми, обробка сигналів, спотворених різного вигляду збуреннями, матричне представлення інформації, що зберігається у великих базах даних.*

Ключові слова: *сегментація, кластеризація, обробка зображень, щільність розподілу,належність*

Введение. Задача кластеризации массивов многомерных наблюдений различной природы, целью которой является нахождение в выборках данных однородных в принятом смысле групп (сегментов, кластеров, классов) наблюдений, является неотъемлемой частью научного направления, именуемого Data Mining [1 – 5], а ее результаты могут быть использованы во множестве приложений. Вместе с тем существует большое число реальных задач, например обработка статических изображений и видеоданных, где обычные методы, в основе которых лежит парадигма самообучения, теряют свою эффективность в силу больших объемов информации, ее «загрязненности» различного рода возмущениями, наличия сегментов сложной формы [6, 7].

Следуя [8], алгоритмы кластеризации и сегментации условно могут быть разбиты на два больших класса: основанные на разбиении и иерархические. Алгоритмы, основанные на разбиении, разделяют группу, содержащую N объектов, описываемых n-мерными векторами признаков $x(k) \in R^n, k = 1, 2, ..., N$, на p кластеров, где p – основной параметр, задаваемый априори. Такие алгоритмы, как правило, стартуют с некоторого достаточно произвольного разбиения, которое в процессе оптимизации некоторой также априори заданной целевой функции, постоянно корректируется с помощью той или иной итерационной процедуры. Основной характеристикой каждого кластера является его центроид–прототип, вокруг которого группируются наблюдения соответствующего класса. Наиболее характерными представителями этого направления являются алгоритмы k-

средних, k-медоидов, нечетких c-средних и т.п. [5]. Вместе с тем, несмотря на множество несомненных достоинств, эти процедуры формируют кластеры строго выпуклой формы, что далеко не всегда имеет место при анализе визуальной информации.

В отличие от этого подхода иерархические алгоритмы, которые, в свою очередь, делятся на агломеративные и «дробящие», автоматически определяют число классов путем слияния отдельных наблюдений в кластеры либо разбиения исходной выборки на подвыборки–кластеры. Здесь, однако, возникает вопрос, когда остановить процесс слияния-дробления, поскольку в пределе алгоритм может сформировать либо один, либо N кластеров. Кроме того, в силу вычислительной громоздкости такие алгоритмы плохо приспособлены для анализа изображений, особенно в ситуациях, когда данные подаются на обработку последовательно.

Альтернативой этим традиционным подходам могут служить методы кластеризации, основанные на плотности распре-деления данных [1, 2], где понятие плотности достаточно близко к плотности распределения, используемой в теории вероятностей и математической статистике. Именно такие методы, позволяют формировать кластеры произвольной формы в условиях, когда данные «зашумлены» возмущениями различной природы. При этом под кластерами здесь понимаются области в пространстве признаков с высокой плотностью распределения данных. Эти области разделены областями с низкой плотностью и именно в этих областях концентрируются возмущения. Таким образом, алгоритм кластеризации, основанный на плотности, в процессе своей работы «выращивает» области с высокой плотностью распределения данных и формирует кластеры произвольной формы, отделяя при этом возмущения и шумы.

Сегментация изображений на основе модифицированного DBSCAN.

Одним из наиболее эффективных и популярных методов кластеризации, основанных на плотности и предназначенных для обработки информации, содержащейся в очень больших базах данных (VLDB – Very Large Data Base), является так называемый DBSCAN (Density Based Spatial Clustering of Applications with Noise), предложенный в [8]. Этот метод характеризуется вычислительной простотой, автоматически определяет количество кластеров произвольной формы, некритичен к действию возмущений.

В основе DBSCAN лежит ряд понятий и определений, основными из которых являются D-досягаемость (Density reachability) и D-связанность (Density connectedness). При этом полагается, что точка $x(k)$ непосредственно D-досягаема из точки $x(q)$, если она удалена (в смысле принятой метрики, обычно евклидовой) на расстояние, не превышающее некоторый порог ε, задаваемый, как правило, априори. Именно величина ε является основным исходным параметром алгоритма.

Таким образом, все точки x, удовлетворяющие неравенству $x–x(q)\leq\varepsilon$, образуют ε-окрестность наблюдения $x(q)$.

При этом, если в ε-окрестности $x(q)$ содержится не менее $MinPts \geq n+1$ точек, считается, что $x(k)$ и $x(q)$ принадлежат одному кластеру. Кроме того, точка $x(k)$ называется D-досягаемой из $x(q)$, если существует последовательность $x(q),\ldots,$ $x(r),\ldots,x(k)$, каждый элемент которой непосредственно D-досягаем своими соседями.

Здесь важно отметить, что отношение D-досягаемости не является симметричным. Например, $x(k)$ может быть граничной точкой кластера и не иметь достаточное количество соседей в своей ε-окрестности. Обычно нахождением таких граничных точек заканчивается формирование каждого конкретного кластера. И наоборот, стартуя из $x(k)$, можно найти точку $x(q)$, имеющую достаточное число соседей. Такие наблю-дения являются внутренними точками кластера. На основе этой асимметрии вводится понятие D-связности, при этом две точки $x(q)$ и $x(k)$ являются D-связными, если они D-досягаемы из $x(r)$. Понятно, что

понятие D-связности симметрично.

Таким образом, в рамках подхода, основанного на плотности, кластер – это множество D-связных точек.

Следуя такой трактовке, процесс кластеризации, стартуя из произвольной точки выборки, находит множество D-связных с нею наблюдений. Далее, продолжая из любой другой не рассмотренной точки, находится следующий кластер и т.д. Все наблюдения, не включенные ни в один кластер, рассматриваются как шумы. Процесс выращивания кластеров идет до тех пор, пока не будут исчерпаны все наблюдения выборки.

Метод DBSCAN в силу своей простоты получил широкое распространение во множестве прикладных задач, в том числе и для сегментации изображений [9, 10], где в соответствии каждому пикселю ставится n-мерный вектор признаков, а все множество таких векторов образует выборку, подлежащую кластеризации. Здесь можно отметить, что в задачах обработки изображений могут возникать ситуации, когда сегменты, содержащие наблюдения с похожими свойствами, могут находиться на достаточно удаленных друг от друга участках изображения, конечно усложняет задачу. Кроме того, DBSCAN может эффективно использоваться и на уровне обобщения информации, когда целое изображение или даже временной сегмент видеоряда рассматриваются в качестве точки признакового пространства, то есть речь, идет о кластеризации коллекций видеоданных с целью организации эффективных средств индексации.

Упростить реализацию и повысить быстродействие в задачах обработки изображений можно, рассматривая это видеокадр не попиксельно, а в виде последовательности прямоугольных фрагмен-тов, а в пределе и полного изображения. Таким самым, задача, рассматриваемая в работе, может быть сформулирована следующим образом. Пусть задана матричная последовательность наблюдений $x(k) = \{x_{i_1,\, i_2}(k)\}$, $i_1 = 1, 2, ..., m$, $i_2 = 1, 2, ..., n$, $k = 1, 2, ..., N$, $x(k) \in R^{m \times n}$. Необходимо с помощью процедуры кластеризации сформировать p однородных сегментов произвольной формы. При этом предполагается, что значение p заранее неизвестно, а сами сегменты могут пересекаться так, что каждое наблюдение может принадлежать нескольким кластерам одновременно.

В процессе решения задачи для каждого матричного наблюдения $x(q)$ формируется ε-окрестность в виде матричной гиперсферы

$$Sp(x - x(q))(x - (q))^T = \varepsilon^2,$$

при этом все объекты $x(k)$, попадающие в эту гиперсферу, являются непосредственно D-досягаемыми из $x(q)$. Аналогично вводится понятие D-связности, а каждый кластер формируется из множества D-связных $(m \times n)$-матриц.

Сам процесс кластеризации происходит совершенно аналогично обычному DBSCAN, однако общее число обрабатываемых объектов при этом существенно сокращается.

Пусть в результате процесса кластеризации некоторой выборки $x(k)$, $k = 1, 2, ..., N$ будет сформировано p кластеров $Cl_1, Cl_2, ..., Cl_l, ..., Cl_p$ и пусть два из них, например Cl_l и Cl_{l+1} в смысле принятой метрики близки друг к другу так, что наблюдение $x(k)$ как граничная точка принадлежит обоим сегментам. В исходном варианте алгоритма [11] точка $x(k)$ будет отнесена к кластеру, который был сформирован первым. Понятно, что такое решение не является наилучшим. В [2] предложена процедура объединения (слияния) двух соседних кластеров в случае, если расстояние между двумя сегментами меньше некоторого значения Δ, т.е. $D(Cl_l, Cl_{l+1}) < \Delta$, где

$$D(Cl_l, Cl_{l+1}) = \min_{x \in Cl_l,\, y \in Cl_{l+1}} D(x, y) =$$

$$= \min_{x \in Cl_l,\, y \in Cl_{l+1}} \sqrt{Sp(x - y)(x - y)^T}\,.$$

На наш взгляд, более естественным представляется рассмотрение данной ситуации с позиций нечеткого подхода к кластеризации [12], при этом наблюдение $x(k)$ с различными уровнями принадлежности может принадлежать сразу нескольким сегментам. В случае двух соседних классов необходимо ввести в рассмотрение их прототипы-центроиды $c(l)$ и $c(l+1)$:

$$c(l) = N_l^{-1} \sum_{\tau=1}^{N_l} x(\tau),$$

$$c(l+1) = N_{l+1}^{-1} \sum_{\tau=1}^{N_l} x(\tau)$$

(здесь N_l и N_{l+1} – число точек, отнесенных к Cl_l и Cl_{l+1} соответственно) и рассчитать соответствующие уровни принадлежности. Обозначая

$$d_{ij} = D^{-1}(x(i), c(j))$$

$$s_{ij} = (Sp(x(i) - c(j))(x(i) - c(j))^T)^{-1/2}$$

для любых допустимых индексов i и j, имеем

$$\mu(x(k), Cl_l) = \frac{d_{kl}}{d_{kl} + d_{kl+1}} = \frac{s_{kl}}{s_{kl} + s_{kl+1}},$$

$$\mu(x(k), Cl_{l+1}) = \frac{d_{kl+1}}{d_{kl} + d_{kl+1}} = \frac{s_{kl+1}}{s_{kl} + s_{kl+1}}.$$

Если же необходимо оценить уровни принадлежности $x(k)$ ко всем кластерам, для расчета может быть использовано соотношение

$$\mu(x(k)), Cl_l) = \frac{D^{-1}(x(k), c(l))}{\sum_{q=1}^{p} D^{-1}(x(k), c(q))} =$$

$$= \frac{(Sp(x(k) - c(l))(x(k) - c(l))^T)^{-\frac{1}{2}}}{\sum_{q=1}^{p} (Sp(x(k) - c(q))(x(k) - c(q))^T)^{-\frac{1}{2}}}.$$

Алгоритм обработки изображений на основе модифицированного DBSCAN представлен на рис. 1.

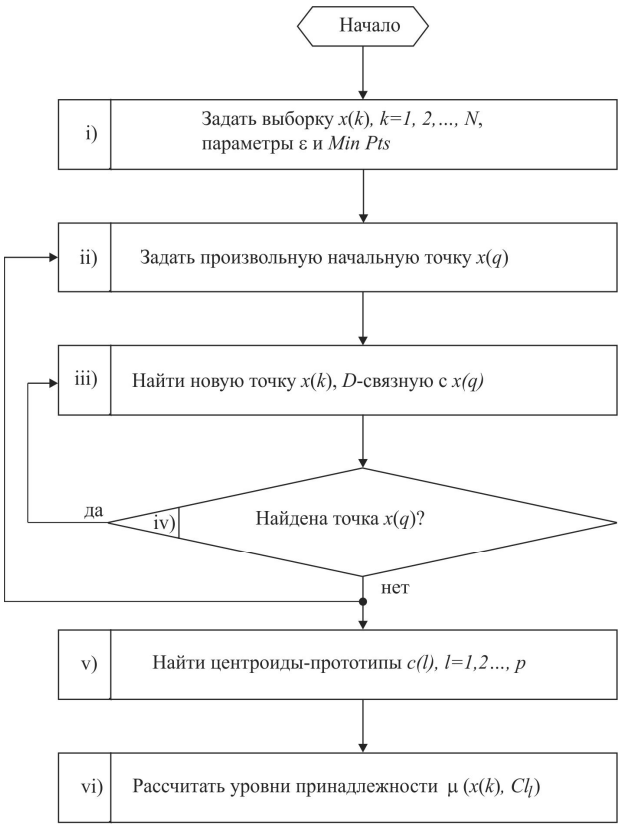

Рис. 1. Алгоритм обработки данных на основе модифицированного DBSCAN

Заключение. В статье предложен метод кластеризации, основанный на плотности распределения информации, содержащейся в больших базах данных. В основе метода лежит модификация DBSCAN, оперирующая матричными наблюдениями и позволяющая оценить уровень принадлежности каждого такого наблюдения к каждому из сформированных кластеров-сегментов. Метод позволяет формировать кластеры произвольной формы, он может обрабатывать данные, искаженные шумами, и по сравнению с прототипом позволяет существенно сократить объем исходной информации.

Список использованной литературы

1. Han J., and Kamber M. Data Mining: Concepts and Techniques, 2-nd ed., (2006), San Francisco: *Morgan Kaufmann*, 800 p.

2. Gan G., Ma C., and Wu J. Data Clustering: Theory, Algorithms, and Applications, (2007), Philadelphia: *SIAM*, 466 p.

3. Abonyi J., and Feil B. Cluster Analysis for Data Mining and System Identification, (2007), Basel: *Birkhäuser*, 303 p.

4. Olson D.L., and Dursun D. Advanced Data Mining Techniques, Berlin, *Springer*, 2008, 180 p.

5. Xu R., and Wunsch D.C. Clustering, (2008), *Hoboken: John Wiley&Sons*, 358 p.

6. Szeliski R. Computer Vision. Algorithms and Applications, (2011), London, *Springer*, 813 p.

7. Bow S.-T. Pattern Recognition and Image Preprocessing, (2002), N.Y., *Basel: Marcel Dekker, Inc.*, 698 p.

8. Kaufman L., and Rousseeuw P.J. Finding Groups in Data: An Introduction to Cluster Analysis, (1990), N.Y., *John Wiley&Sons*, 342 p.

9. Дуда Р. Распознавание образов и анализ сцен / Р. Дуда, П. Харт. – М. : Мир, 1976. – 512 с.

10. Форсайт Д. Компьютерное зрение: современный подход / Д. Форсайт, Ж. Понс, – М. : Издательский дом «Вильямс», 2004. – 928 с.

11. Ester M., Kriegel H.-P., Sander J., and Xu X. A density-based Algorithm for Discovering Clusters in Large Spatial Database with Noise, (1996), *Proc. Int. Conf. on Knowledge Discovery in Databases and Data Mining, Portlend, Oregon: AAAIO Press*, pp. 226 – 331.

12. Bezdek J.C., Keller J., Krisnapuram R., and Pal N.R. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, (2005), N.Y., *Springer Science+Business Media, Inc.*, 776 p.

Получено 07.04.2014

References

1. Han J., and Kamber M., (2006), Data Mining: Concepts and Techniques, *Morgan Kaufmann*, San Francisco, 800 p.

2. Gan G., Ma C., and Wu J., (2007), Data Clustering: Theory, Algorithms, and Applications, *SIAM*, Philadelphia, 466 p.

3. Abonyi J., and Feil B., (2007) Cluster Analysis for Data Mining and System Identification, Birkhäuser, *Basel*, 303 p.

4. Olson D.L., and Dursun D., (2008), Advanced Data Mining Techniques, *Springe*, Berlin. – 180 p.

5. Xu R., and Wunsch D.C., (2008), Clustering, John Wiley&Sons, *Hoboken*, 358 p.

6. Szeliski R., (2011), Computer Vision. Algorithms and Applications, *Springer*, London, 813 p.

7. Bow S.-T., (2002), Pattern Recognition and Image Preprocessing, *Marcel Dekker, Inc.*, New-York. – 698 p.

8. Kaufman L., and Rousseeuw P.J., (1990), Finding Groups in Data: An Introduction to Cluster Analysis, *John Wiley&Sons*, New-York, 342 p.

9. Duda R., and Hart P., (1976), Raspoznavaniyeobrazov I analizscen [Pattern Recognition and Scenes Analysis], *Mir*, Moscow, Russian Federation, 512 p. (In Russian).

10. Forsyte D., and Pons J. Kompyuternoezreniye: sovremennyypodhod [Computer Vision: Modern Approach], (2004), *Izdatelskiydom "Willians"*, Moscow, Russian Federation, 928 p. (In Russian).

11 Ester M., Kriegel H.-P., Sander J., and Xu X., (1996), A Density-based Algorithm for Discovering Clusters in Large Spatial Database with Noise, *Int. Conf. on Knowledge Discovery in Databases and Data Mining, AAAIO Press, Portlend, Oregon*, pp. 226 – 331.

12 Bezdek J.C., and Keller J., Krisnapuram R., and Pal N.R., (2005), Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, *Springer Science+Business Media, Inc.*, New-York, 776 p.

Богучарский Сергей Иванович, аспирант каф. информатики, Харьковского нац. ун-та радиоэлектроники, пр. Ленина, 14, Харьков, Украина, 61166, тел. +38067-483-47-18. E-mail: sbogucharskiy@rambler.ru

Машталир Владимир Петрович, декан ф-та компьютерных наук, Харьковского нац. ун-та радиоэлектроники E-mail: mashtalir@kture.kharkov.ua